IN23A-1413

Science & Technology Facilities Council

Science Support: The Building Blocks of Active Data Curation

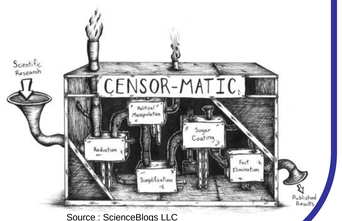
Anabelle Guillory & The Centre For Environmental Data Archival (CEDA) Team
Centre for Environmental Data Archival, STFC, Harwell Oxford, Oxfordshire, UK OX11 OQX
Anabelle.Guillory@stfc.ac.uk

The Data Deluge

Today, a single dataset can range from several gigabytes (research campaigns) to a data deluge encompassing multiple petabytes (e.g. UPSCALE) and soon exabytes. The challenge scientists face is not only finding facilities to analyse "Big Data" but also ensuring the long term research value of their data by making them available for further high quality research...This is at the heart of Science Support at the Centre for Environmental Data Archival (CEDA).

"Good research needs good data", Data Curation Centre

Conclusions and knowledge are only as good as the data they're based on. It's up to us as scientists to care for the data we've got, (whatever size it is) and ensure that the story of what we did to the data is transparent so that we can use the data again and so others will trust our results. Science Support activities exist to ensure that unique and irreplaceable project data outputs have a long term future.

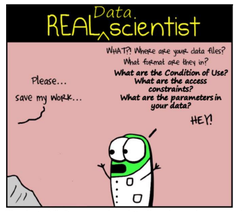


Source: ScienceBlogs LLC

Scoping Data & Data Management Plan

Effective digital curation relies upon sound planning. At the core of Science Support activities is the Data Management Plan (DMP) which sets out a coherent approach to data issues pertaining to the data generating project. ***"Plans typically state what data will be created and how, and outline the plans for sharing and preservation, noting what is appropriate given the nature of the data and any restrictions that may need to be applied"*** – Data Curation Centre website.

DMP or Outlined DMP are now often required as part of a funder's award application process. A DMP addresses the associated project data management issues and assists the project members to make appropriate data decisions.



Thanks to Nik Papageorgiou : http://upmic.wordpress.com/

Following the completion of a **Scoping Questionnaire**, the PI and the Data Centre develop, write and mutually agree on the DMP which includes:

- Delivery schedules
- Conditions of Use/Licensing
- Responsibilities of data providers & CEDA
- Project specific requirements
- 3rd party data requirements
- Collaborative group workspace

Data Analysis Platforms & Storage Facilities at CEDA

With decades worth of satellite data and high volume complex climate datasets, CEDA offers a **High Performance Storage Environment,** referred to as **JASMIN.**

In line with the Data Curation Lifecycle Model, JASMIN not only provides multi-petabyte storage and back-up facilities but it also allows users to manipulate the data in-situ (e.g. with the deployment of virtual machines running custom scientific software co-located with data), **generating processed data outputs ready for curation within the data centre!**

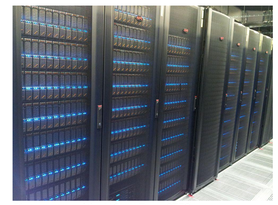


Photo courtesy of the JASMIN Team, STFC

Providing access to JASMIN capabilities (e.g. data intensive computing, storage, collaborative workspaces, etc…) is a CEDA Science Support activity which is becoming increasingly more important as users around the world find data analysis ever more challenging at their home institution.

For more information on JASMIN, go to http://jasmin.ac.uk/

Archiving Metadata and Data Publication

In agreement with the DMP, the data centre sets up **data ingest routes,** back-up storage and provides guidance and assistance with **format issues** and **metadata conventions**, **file naming and archive structure**, as well as a **helpdesk** to handle queries from data users and providers.

created by Jørgen Stamp http://www.jstamp.dk

M**etadata – data about data** is critical for future re-use of data. It is the role of the data centre to capture supporting documentation (formats, calibration information, flight logs, model details etc.) ensuring it is correct and useful and to put it into structures where it can be used to provide tools and services. At CEDA, all metadata is made available through the NERC Data Catalogue Service which increases the **visibility and discoverability of datasets**, thereby increasing their impact.

To address the challenge of "Big Metadata", **automatic metadata collection** is currently being explored as part of the ExArch project.

CEDA provides and encourages a mechanism for **data citation and publication** (minting DOIs (Digital Object Identifiers)) which allows data producers **to receive full academic credit** for their efforts in producing and documenting the data.

From Supporting Small Research Campaigns…

There is no minimum dataset size requirement when it comes to data curation. Data from small research measurement campaigns are just as valuable as large generated output (e.g. instrument calibration, mesoscale research).

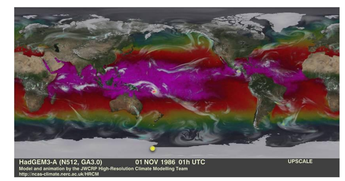
The NERC/Met Office BAe 146-301 large Atmospheric Research Aircraft, operated by FAAM (Facility for Airborne Atmospheric Measurements) is deployed anywhere in the world on science missions measuring atmospheric properties and occasionally in support of civil contingency (e.g. Eyjafjallajökull volcanic eruption).

Above & Left, courtesy of FAAM.
Below, courtesy of the JASMIN Team, STFC.

All FAAM Campaign data are archived at CEDA-BADC.
FAAM depend on access to CEDA-BADC Services (FTP server, archive, collaborative workspace, trajectory service, file format and metadata information) whenever and wherever they are.

…To Super-size Research Projects…

The UPSCALE project (UK on PRACE: weather-resolving Simulations of Climate for globAL Environmental risk) is the largest ever PRACE (Partnership for Advanced Computing in Europe) computational project, led by the UK**,** and dependent on CEDA-BADC to provide the data links and data analysis environment.



Picture courtesy of P-L Vidale and R. Schiemann, NCAS

"We would never have been able to store, nor analyse, that volume of data, without the existence of the [JASMIN] service..." - UPSCALE spoke-scientist

CEDA Support for the project included: virtual machines for NCAS and Met Office, 380Tb of group workspace storage and temporary backup (~200Tb), high-bandwidth allowed retrieval of 250Tb from Germany in 1 year, over 100Tb pulled back to Met Office archive.

UPSCALE has so far generated 0.5PBytes of data which will be widely studied over the next decade!

British Atmospheric Data Centre
NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE
NATURAL ENVIRONMENT RESEARCH COUNCIL

Centre for Environmental Data Archival
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESERACH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL